Critical Reviews in Toxicology, 2011; 41(4): 321–338

ISSN 1040-8444 print/ISSN 1547-6898 online
DOI: 10.3109/10408444.2010.541222

informa
healthcare

REVIEW ARTICLE

Historical perspective on the use of animal bioassays to predict carcinogenicity: Evolution in design and recognition of utility

L. A. Beyer,[1] B. D. Beck,[1] and T. A. Lewandowski[2,3]

[1]Gradient, Cambridge, Massachusetts, USA, [2]Gradient, Seattle, Washington, USA, and [3]Department of Health and Nutrition Sciences, Brooklyn College, The City University of New York, Brooklyn, New York, USA

Abstract

The animal testing protocols used today to evaluate the carcinogenicity of chemicals are very different from those used in the earlier part of the 20th century. To explore how cancer bioassays have changed over time, we surveyed the literature discussing test design and interpretation from the 1930s to the present. We also analyzed compendia of bioassays published by the US Public Health Service (US PHS) from 1938 to 1978, and evaluated the data to understand the evolution of testing methodology (e.g., animals used, test duration) and the types of chemicals being studied. The cancer bioassay evolved in several stages. At the beginning of the 20th century, animal bioassays were primarily used to re-create known human diseases, whereas in the 1940s to 1960s, animal bioassays were largely used to evaluate the safety of chemicals in foods, drugs, and cosmetics. Beginning in the late 1960s and 1970s, chemicals primarily associated with occupational or environmental exposures were also evaluated. Testing strategies now emphasize a suite of tests including multiple in vitro tests and both short-term and long-term animal tests. The objectives of testing are broader, too, with test goals encompassing information regarding mode of action and other parameters aimed at evaluating potential species differences (e.g., in toxicokinetics) and their relevance for evaluating human risks. It is important to consider this evolution when evaluating the testing methodology and scientific conclusions in earlier eras. As toxicology continues to develop, testing methods will continue to change in concert with increased knowledge and understanding.

Keywords: Animal bioassays; animal testing; cancer bioassays; chronic animal testing; historical analysis; US public health service

Contents

Address for Correspondence: Thomas A. Lewandowski, Ph.D., Gradient, 600 Stewart Street, Seattle WA 98101, USA. E-mail: tlewandowski@gradientcorp.com

(Received 18 January 2010; revised 23 October 2010; accepted 15 November 2010)

1900s to the 1950s: Early toxicological testing

Reports of animal experiments designed to investigate the induction of tumors by chemical agents appeared as early as 1915, when Yamagiwa and Ichikawa published the results of studies of the carcinogenicity of coal tar extracts using skin painting studies in rabbits (Yamagiwa and Ichikawa, 1915). In subsequent decades, animal studies were also conducted to explore the health effects of coal tars (Tsutsui, 1918; Yamagiwa and Ichikawa, 1918; Murray, 1922), benzene (Selling, 1916), β-naphthylamine (Hueper et al., 1938), and other cancer-causing agents.

When reviewing the commentary provided in these publications, it is clear that these animal experiments were not generally used as screening tools for assessing the potential health effects of chemicals (whether cancer or other effects), but rather for studying the disease process itself. Safety evaluations of acute exposures were sometimes conducted, often with lethality as the primary endpoint (Patty et al., 1930), but the idea of using animal testing to screen for the potential to cause chronic effects such as cancer was not discussed. Predictive testing for foods and drugs became a higher priority after the well-known poisoning episode in 1937 in which the diethylene glycol used as a diluent in Elixir of Sulfanilamide caused the deaths of 105 individuals, many of whom were children (Ballentine, 1981; Wax, 1995). This scandal brought to light the need to verify the safety of chemicals in foods and drugs prior to their entry into the marketplace (Wax, 1995). In response to public pressure, the Food, Drug, and Cosmetic Act of 1938, which had previously languished for several years in Congress, was quickly passed (Wax, 1995). The new law strengthened provisions of the earlier Pure Food and Drug Act and added provisions for setting tolerances for "poisonous or deleterious" substances in food and for requiring pharmaceutical manufacturers to submit data on drug safety to the US Food and Drug Administration (FDA) for approval (FDA, 2010).

Beginning in 1943, FDA scientists published a series of journal articles and reports discussing the use of long-term tests for safety assessment of food and drugs (e.g., Nelson et al., 1943, 1945; Woodard and Calvery, 1943; Fitzhugh and Nelson, 1946; Lehman, 1951; Lehman et al., 1949, 1955; FDA, 1959). Because the presence of potential toxicants in drugs, and especially in foods, could result in potentially widespread exposures with effects that would be difficult to observe directly, FDA scientists relied on animal studies to evaluate new food and drug ingredients. FDA′s head of pharmacology, Arnold Lehman, in his article "Proof of Safety, Some Interpretations" discussed the importance of testing chemicals added to foods, drugs, or cosmetics, noting that food additives require the most thorough appraisal, in part due to the potential for widespread exposure in the general population (Lehman, 1951). In his article, Lehman mentioned long-term bioassays of seven chemicals (i.e., glycols, ergot, rancid lard, selenium, thiourea, and the food sweeteners 5-nitro-2-propoxyaniline and dulcin), all of which were found in, or added to, foods or pharmaceuticals.

Under the direction of Lehman, FDA scientists pioneered the use of animal studies to predict potential chemical hazards (including carcinogenicity) in humans, and laid the early foundation for the use of animal data in human health risk assessment. FDA′s first guidance to industry, "Procedures for the Appraisal of the Toxicity of Chemicals in Food," was published in 1949 (FDA, 2005). These procedures were largely adopted by laboratories conducting investigations of chemical food additives (Lehman et al., 1955). In 1955, FDA updated and renamed the guidance document to reflect the new emphasis on drugs and cosmetics: "Procedures for the Appraisal of the Toxicity of Chemicals in Food, Drugs and Cosmetics" (Lehman et al., 1955). Lehman et al. noted that conducting animal studies to set a safe dose in humans had been successful, but perhaps:

> the major lesson that has been learned from the review of animal data on several thousand compounds is that in order successfully to evaluate potential toxicity in man we must have some understanding of each of the many facets of the action of the chemical. Acute- and chronic-toxicity experiments alone are not enough. We must have data also on histopathology, hematology, pharmacodynamics, irritation, and sensitization. Studies on absorption, distribution, excretion and duration of sojourn in the body are also helpful.

The National Cancer Institute′s (NCI′s) Jonathan Hartwell also commented on the amount of data that would need to be acquired in order to reliably extrapolate test results in animals to humans:

> Another pitfall is the attempt to carry over, without reservation, to man, conclusions based on animal experiments. We do not know whether man is more or less susceptible than mice to particular carcinogens. Some animal species, such as the rat, rabbit and dog, are much more resistant to certain chemical carcinogens than is the mouse, and vice versa, while in the monkey none of the powerful carcinogens has been shown to produce tumors. It would, therefore, be dangerous to conclude that man is resistant or susceptible to a given carcinogen merely on the basis of experiments with a single species of laboratory animal since we also know of several

chemical agents which, as occupational hazards, are responsible for some human cancer and which have so far yielded negative results in laboratory animals. In the light of species specificity, a given compound, whether or not it is carcinogenic to animals, may or may not be capable of producing cancer in man. In view of the large number and variety of carcinogenic compounds, we must regard this multiplicity of newly developed and yet to be developed drugs, dyes, food additives and other chemical compounds with which we come in personal contact as possible cancer hazards, and we must reserve judgment on specific compounds until they are adequately investigated. (Hartwell, 1951)

Thus, there was recognition among those scientists at the forefront of toxicity testing that the amount of data required to fully understand a chemical's toxicity (including its carcinogenicity) was substantial, particularly with respect to extrapolation to humans, and not likely to be completely met by the limited studies that were typically being conducted.

The 1959 edition of the FDA guidance contained a special chapter on "carcinogenicity screening," which provided advice on such issues as dose route selection, species selection, and genetic composition. It also discussed the difficulties of detecting the carcinogenicity of less potent carcinogens, noting that it "is not practicable to formulate a set of rules for carcinogenicity testing that would apply under all circumstances because testing requirements will vary with the carcinogenic potency and extent of the proposed use of the compound" (FDA, 1959). To evaluate carcinogenicity and other forms of chronic toxicity in food additives and pesticides, FDA recommended starting with lifetime studies in two species: the rat and a non-rodent such as the dog or non-human primate (FDA, 1959). In FDA's protocol, both rats and dogs were tested for 2 years and divided into four dose groups: control, probable no effect level, mid-dose group, and high-dose group (chosen so as to be an unambiguous adverse effect level). All animals were necropsied for gross pathology and selected organs were weighed and preserved for histopathological study (FDA, 1959).

Similar concern in Europe over the safety of food additives resulted in the publication of *Procedures for the Testing of International Food Additives to Establish Their Safety for Use* in 1958 by the World Health Organization (JECFA, 1958). This document noted the early recognition that high doses ("far in excess of those recommended for human consumption") were an acceptable approach in lieu of using the hundreds of animals that would be required at lower doses to achieve statistical significance (JECFA, 1958, pp. 8–9), and that "the use of high dosage levels of the test substance and the spread of the investigations over a number of different species make it reasonable to extrapolate the data to man." Procedures were provided for conducting acceptable acute, subacute, and chronic toxicity studies (e.g., 25 rats/sex/dose with two dose levels and one control group). These guidelines were updated in 1977 (as cited in IARC, 1979).

Through the 1950s and 1960s, chemicals subjected to the most intense toxicological evaluation were generally drugs, direct food additives (such as chemicals added to margarine to simulate butter color and flavor), indirect food additives (e.g., chemicals that could potentially migrate from food packaging materials), and pesticide residues (Junod, 1999; FDA, 2005; NRC, 2007). FDA's interest in pesticides and pesticide residues in food is indicated by a comprehensive review published by Agency scientists in 1950 (Rohwer et al., 1950).

1960s and 1970s: Increasing reliance on bioassays for evaluating risks of occupational and environmental chemicals

As a result of mounting public concerns in the 1960s and 1970s, increased attention was devoted to conducting chronic bioassays with chemicals found in the workplace and the general environment. Prior to this time, the occupational health risks of certain classes of chemicals (e.g., pesticides) had been evaluated in animal studies (Hall, 1951; Fairhall, 1952), but such testing was not routine nor was it conducted for most industrial chemicals. In Rachel Carson's widely read *Silent Spring* (Carson, 1962), Carson severely criticized both industry and government for permitting the widespread use of chemicals (chiefly pesticides) without developing a full understanding of their potential ecological and human health effects. From within the government, Wilhelm Hueper at the US National Cancer Institute (NCI) campaigned vigorously for better assessment of environmental and occupational hazards (e.g., Hueper, 1957), which as noted by the American Conference of Governmental Industrial Hygienists (ACGIH) (in the context of a discussion on setting threshold limit values (TLVs) for the workplace in the 1940s), had been hampered due to the lack of sufficient toxicological data (Frederick, 1984).

Another major impetus to developing the cancer bioassay protocol occurred when NCI recognized, in 1961, "that there was a need for more systematic investigation on chemical carcinogenesis in animals" (Weisburger, 1983). In part this reflected concerns about potential chemical carcinogens in food and the need to satisfy the requirements of the Delaney clause (amending the Food Drug and Cosmetic Act in 1958), which prohibited the introduction of chemicals into the food supply which "induce cancer in man or animal" (Weisburger, 1994).

The time was ripe for an overhaul of the bioassay concept prior to dramatically expanding the use of such tests. There were many frustrations with bioassays, as discussed by Hackmann (1958), who stated that "Today, after having tested hundreds of chemical substances, we often still find it difficult to decide whether a substance is dangerous to man or not."

Under the leadership of Michael Shimkin, two NCI researchers, Elizabeth and John Weisburger, began in 1961 to develop a systematic approach for testing chemicals at NCI (Weisburger, 1999), building on the testing

recommendations previously set forth by the scientists at FDA. As part of this effort, they surveyed the problems with bioassays conducted up to that time (Weisburger and Weisburger, 1967):

In the past, …studies were deficient in one or more aspects: (1) poorly defined strain of the species used, (2) inadequate details on amounts of compound administered, (3) insufficient data on numbers and starting age of animals, length and frequency of treatment, latent period to first tumor appearance and to average tumor appearance, (4) lack of description of dietary conditions, intake, weight gain, etc., (5) deficient histopathology, and (6) no, or imperfect, control series.

The Weisburgers emphasized the use of what became the standard cancer bioassay protocol, using both rats and mice, oral gavage dosing, and terminating the experiment at 92 weeks of age in mice and 104 weeks of age in rats (Weisburger and Weisburger, 1967; Weisburger, 1999). Specific strains of rats and mice (F344 rats and B6C3F1 mice) were selected for their low rate of spontaneous tumors and apparent sensitivity to chemical carcinogens (Weisburger, 1999). This systematic method for assessing carcinogenicity of chemicals in rodents had tremendous impact on subsequent government testing activity (NTP, 1984). Although the Weisburgers′ recommended testing protocols were informed by the experimental studies previously conducted by FDA scientists (Weisburger, 1983), they brought a degree of standardization to the animal bioassay that had been sorely lacking.

Beginning in the late 1960s, chronic animal testing became increasingly conducted on large numbers of chemicals simultaneously. For example, in 1968 researchers published an article describing the results of studies involving 38 chemicals that were known or suspected animal carcinogens, including derivatives of aromatic amines, nitrosamines, quinolines, and purine antimetabolites (Hadidian et al., 1968). The goal of this testing was "to explore current methodology to develop a technique suitable for mass screening…in view of the growing need to gain information on ever-increasing environmental hazards." The researchers chose the oral route, because it was a "major pathway of exposure to man" and used four to six dose levels, as recommended by FDA. An 8-week subchronic study was performed to identify the lethal dose. The long-term study was then conducted with three male and three female rodents per dose; the animals were dosed five times per week for a total of 260 doses in 1 year. The protocol then allowed another 6 months for reversible damage to be repaired and neoplastic lesions to be intensified. The authors commented on the dilemma (which persists today) of balancing the number of dose groups with the number of animals per group (given a fixed number of animals). They suggested that, for chemicals suspected of being weak carcinogens, more rats be tested at each dose to increase the power to detect a response.

Innes and colleagues published the results of long-term bioassays performed on 120 different compounds for NCI (Innes et al., 1969). Mice of two different strains were dosed via gavage starting at 5 days of age and then via the diet from age 4 weeks until terminal sacrifice at 18 months. According to the authors, the chemicals evaluated were chosen based on chemical structure, widespread use, and evidence of toxicity in the literature. We examined the list of chemicals tested by Innes et al. (1969) and supplemented their chemical use data with additional information in order to understand the human health relevance of the chemicals selected for testing by NCI during this period.

As shown in Figure 1, 88% of chemicals tested were pesticides (88% is the combination of fungicides [42%], herbicides [25%], insecticides [18%], and rodenticides [3%]), whereas an additional 9% were primarily additives used in rubber manufacturing, an industry in which elevated carcinogenic risks in workers had been known since the 1940s (Case and Hosker, 1954). FDA tolerances based on contact of rubber parts with food items existed as early as the 1970s for most of these chemicals, suggesting that these tests were part of the tolerance setting process. The remaining 3% of the chemicals were used in pharmaceutical production, plastic production, and heat transfer fluids.

These data suggest that in the mid- to late 1960s, the majority of carcinogenic testing in the federal government remained focused on direct and indirect exposures to chemicals in food. However, more research was devoted to understanding chemical hazards in the workplace after the passage of the Occupational Safety and Health Act in 1970, which led to the establishment of the Occupational Safety and Health Administration (OSHA) and its research counterpart, the National Institute of Occupational Safety and Health (NIOSH, 1996).

It was not until the late 1960s, when the standardization of test designs began, that there was general agreement (with some cases of dissent) that animal testing was useful in predicting human health risks. The scientific consensus appeared to be that many of the problems of the early era had been addressed and that chronic toxicity testing in animals, although not an exact model of human disease, was the best option available. In 1972, Dr. Umberto Saffioti, Assistant Director of NCI′s Carcinogenesis Program from 1968 to 1976, stated that "the only prudent course of action is to assume that chemicals which are carcinogenic in animals could also be so in man" (Breslow et al., 1978). A few years later, he noted that "the detection of carcinogenicity in animal experiments has to be regarded as an essential warning signal for human cancer prevention" (Keplinger et al., 1975).

The expanded use of the bioassay protocol received another boost when NCI funding for testing increased dramatically in the early 1970s, largely due to the passage of the National Cancer Act in 1971. NCI′s budget was nearly doubled and the agency had considerable independence in setting its research agenda (NCI, 2007a). Animal testing to both identify new carcinogens and understand carcinogenic mechanisms was conducted

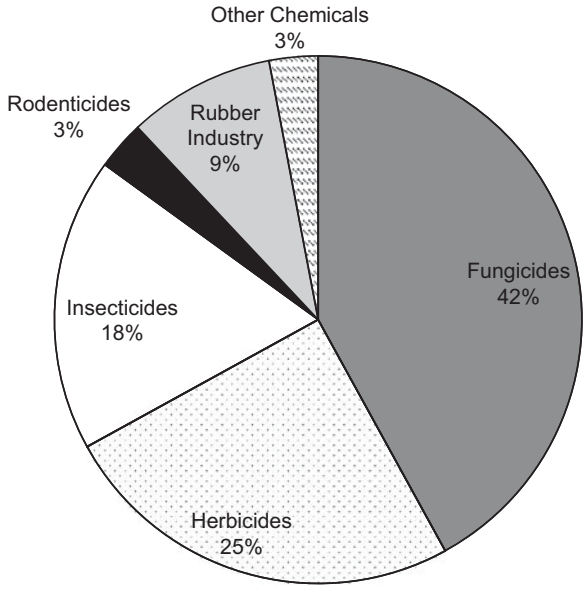

Figure 1. Classes of chemicals studied by the National Cancer Institute in the late 1960s, as reported by Innes et al., (1969).

in the agency′s own laboratories or via agreements with contract laboratories (e.g., Litton Bionetics, Hazleton Laboratories [Kalberer and Newell, 1979])].

The additional testing under the auspices of NCI led to the recognition that standardization of protocols for large-scale carcinogen bioassays was needed to increase scientific acceptability. Such protocols facilitated comparison of results across chemicals, and provided new data regarding comparative toxicological potency. This realization led to a Workshop on Carcinogen Bioassay Protocols, held in November 1973, which resulted in the development of chronic toxicity and carcinogenicity guidelines (Page, 1977) as well as the development of guidelines for conducting carcinogen bioassays in small rodents, as articulated by Sontag et al. (1976) of NCI. These detailed guidelines, which were also used by the World Health Organization (WHO) as early as 1979 (preamble to IARC, 1979), represent the beginning of the modern standardized approach to carcinogenicity bioassays as it is now understood.

Another event that radically changed the perception of the predictive value of animal testing for occupational health hazards was the discovery that vinyl chloride (VC) could produce hepatic angiosarcoma in both animals and humans. The Italian researcher, Cesare Maltoni, presented findings at a conference in Bologna, Italy, in 1973 showing that chronic inhalation exposures of rats, mice, and hamsters to 250 ppm VC produced hepatic angiosarcomas. The results were subsequently published in 1974 in conference proceedings and in the peer-reviewed literature (Maltoni and Lefemine, 1974a, 1974b). In January 1974, a cluster of angiosarcoma cases in VC workers was reported at the BF Goodrich Company in Louisville, Kentucky (Creech and Johnson, 1974). This unusual co-occurrence of findings illustrated that animal tests had predictive potential. As Maltoni himself noted:

Bioassays on the long term effects of VC were not available until 1970. This situation reflects the limited consideration which has been, up to the present, given to experimental bioassays in predicting the oncogenic effects on man of environmental and occupational agents… What is now happening with vinyl chloride and the story of vinyl chloride carcinogenicity should bring about greater consideration of experimental bioassays and finally induce a new turn in the field of occupational and environmental carcinogenesis.

Thus, experimental prediction of the pathogenic potential of occupational and environmental substances before the substances were produced and released on a large scale into the human environment was being seen as an emerging alternative to the typical practice of waiting for evidence in exposed human populations (Maltoni and Lefemine, 1975). Over the next decade, Maltoni and coworkers at the Ramazzini Foundation in Bologna, Italy, evaluated the carcinogenicity of a large number of industrial chemicals (Soffriti et al., 2002).

In the 1970s, the use of animal study data to quantitatively evaluate potential carcinogenicity in humans became established policy worldwide. In 1971, the International Agency for Research on Cancer (IARC), an intergovernmental agency forming part of the WHO of the United Nations, initiated a program to evaluate the carcinogenic risk of chemicals to humans, and a year later began publishing monographs on the "Evaluation of Carcinogenic Risk of Chemicals to Man." The goal of the monographs was the compilation, by experts, of a compendium on carcinogenic chemicals, including the evaluation and documentation of the biological activity of "practical importance to public health" (IARC, 1972, p. 8). The monographs provided a critical assessment of the animal data to help authorities make decisions concerning preventive measures or legislation. For the most part, animal studies were included in the monographs only if they met certain criteria (e.g., at least 25 animals of each sex evaluated for at least 18 months in the case of mice and 2 years in the case of rats; adequate pathological examination) (IARC, 1972, p. 13).

IARC continued to refine its methodology. For example in 1979, more detail was provided defining adequate studies, and categories of evidence for carcinogenicity in animals were established: "limited evidence" and "sufficient evidence" of carcinogenicity. In addition, the use of short-term tests was discussed (IARC, 1972, p. 15).

In the United States, FDA was the first federal agency to use quantitative risk assessment techniques in interpreting animal bioassay data. In 1973, FDA proposed a mathematical procedure, first described in the scientific literature by Mantel and Bryan (1961), for calculating maximum acceptable carcinogen intake levels in food. This risk assessment procedure involved using a conservative log-probit extrapolation (a threshold model), combined with a definition of insignificant risk resulting in what Mantel and Bryan labeled a "virtually safe dose" for carcinogens, which FDA later termed a "safe"

dose (Rodricks, 1988). The United States Environmental Protection Agency (US EPA), following FDA′s lead, developed guidelines for assessing risk from exposures to chemicals in the environment (e.g., air, water, and soil) and published its first set of interim guidelines on carcinogen risk assessment in 1976 (US EPA, 1976).

Consistent with the notion of conducting dose-response extrapolation to relate animal test data to potential human exposures, the National Center for Toxicological Research (NCTR) commenced in 1974 an experiment intended to obtain very robust data on tumorigenic potential at relatively low levels of exposure (FDA, 2009). As noted above, toxicologists had long been concerned with the need to maximize the number of animals per dose group so as to achieve statistical power to distinguish a chemically related response from background. The NCTR study was designed to detect a 1% increase in tumor response above background (hence its designation as the ED01 study). It did so by using over 24,000 animals and the known genotoxic carcinogen, 2 acetylaminofluorene (2-AAF). Yet far from providing a definitive answer concerning low-dose carcinogenicity, it gave mixed results, with an indication of a threshold for bladder tumors but no threshold for liver tumors (Gaylor, 1980).[1]

In 1978, the National Toxicology Program (NTP) was established within the US government as a cooperative effort to bring together the government′s major toxicological research and testing activities to evaluate the human health effects of chemical agents in our environment (NTP, 2004). The NTP assumed the bioassay program previously conducted by the NCI and also coordinated the work of independent toxicology testing programs associated with other federal agencies (e.g., National Institute of Environmental Health Sciences [NIEHS], FDA, and NIOSH), universities, and private groups. The review and selection of substances and issues nominated for study were designed as a multistep process in which a broad range of concerns are addressed through the participation of the NIEHS, other federal agencies, the NTP Board of Scientific Counselors, the NTP Executive Committee, and the public. Another key activity of NTP has been the periodic publication of the "Report on Carcinogens," which synthesizes the available scientific research on the carcinogenicity of specific chemicals, develops a consensus position about a given chemical′s carcinogenic potential, and identifies other chemicals that have been recommended for future carcinogenicity studies (NTP, 2006).

Despite the widespread acceptance of animal testing, there were still a number of criticisms of long-term carcinogenicity tests: the use of administration routes different from those typical of human exposure, doses much higher than those experienced by humans, and an experimental model that was too sensitive (e.g., liver tumors in mice) (Tomatis, 1979). Nonetheless, good correlation was observed between human and animal data such that the target organ of at least one of the experimental animal species was the same as that in humans, regardless of the route of exposure (Tomatis, 1979).

1980s: Refinement of bioassays and use of bioassay data in risk assessment

In the 1980s, animal bioassays were widely used in assessing carcinogenic risk to humans even as more emphasis was placed on mutagenicity and other short-term tests. In an effort to speed up and simplify the hazard identification process, researchers attempted to find a short-term screening test (e.g., in bacteria or fungi) that could predict the findings of whole-animal studies (Weisburger, 1999). Assays such as the bacterial Ames test (McCann et al., 1975; Maron and Ames, 1983) became increasingly recommended as part of carcinogenicity testing batteries (e.g., US EPA, 1986a, 1986b; Weisburger, 1999). In addition, testing protocols became more rigorous with new requirements including the use of Good Laboratory Practice (GLP) and more comprehensive and standardized histopathological examination.

In 1980, IARC published a collection of critical reports on the use of long- and short-term assays for the detection of chemical carcinogens, titled Long Term and Short Term Assays for Carcinogens: A Critical Appraisal (Supplement 2 to the IARC Monographs, IARC, 1980). As reported by the director of IARC, the major objective was to encourage scientists to meet basic requirements when conducting and reporting the results of the studies, because "the availability of results from adequately conducted and critically analyzed experiments is essential to strengthening measures for the primary prevention of cancer" (Tomatis, 1986). Similarly, in the early 1980s, NTP added a group of five in vitro short-term cellular and genetic toxicology assays for measuring gene mutations (bacterial and mammalian cells), chromosome damage, mammalian cell transformation, and DNA damage/repair to the prechronic phase of the toxicology and carcinogenesis bioassay process (Rall, 1984, p. 289).

In 1986, IARC published an updated and revised Supplement 2 as Publication No. 83 (IARC, 1986) although the document retained a similar title. The report included lengthy sections on conducting bioassays for carcinogenicity in animals with particular attention given to the role of pharmacokinetic data in the design and evaluation of such assays, the role of early preneoplastic lesions in carcinogenesis, and the relative contributions that carcinogens may make to the various stages of the carcinogenesis process. However, most of the report discussed

[1]Similar difficulties in determining the exact shape of the dose-response curve at low doses were also encountered in a later large study (4800 rats total) involving N-nitrosodiethylamine or N-nitrosodimethylamine (Peto et al., 1991). Although such studies have not resolved the question of low-dose linearity for carcinogens, they have provided extensive data sets that have been used to formulate alternative models for assessing carcinogenic risk (SOT, 1981; Gaylor and Aylward, 2004).[2]Recall that dermal studies of multicyclic aromatic compounds (e.g., skin painting studies) were not included in our analysis.

short-term bioassays to predict carcinogenicity, noting that although the correlation between mutagenic and carcinogenic effects was "far from perfect," the intelligent use of short-term testing was essential. The main developments in recent years had been a better understanding of the endpoints measured, better validation of the capacity of the short-term bioassays to detect carcinogens, and characterization of new assays using prokaryotic and eukaryotic cells (Tomatis, 1986).

NTP convened an Ad Hoc Panel on Chemical Carcinogenesis Testing and Evaluation that resulted in a lengthy publication describing NTP's testing protocols, which required a comprehensive histopathological examination of 31 to 33 organs/tissues (NTP, 1984, p. 194). This document noted the fact that GLP, which was legislated in the United States and published in the *Federal Register* in 1978 (FDA, 1978, p. 59986), was critical to follow to ensure study integrity and had exerted substantial influence on the conduct and reporting of animal studies carried out for regulatory purposes (NTP, 1984, pp. 12, 13). GLP guidelines had been articulated as early as 1976 by FDA and were similar to NCI Guidelines for Carcinogen Testing instituted on January 1, 1974 (Page, 1977). Similarly, diagnostic guidelines and standardized terminology for the evaluation of animal tumors were also developed and adopted within this timeframe (Ward et al., 1978; Boorman et al., 1985).

At about the same time, GLP was also codified and adopted in Europe by the Organisation for Economic Co-operation and Development (OECD), which developed the principles of GLP to promote the quality and validity of test data used for determining the safety of chemicals and chemical products. The principles, elaborated by an Expert Group on GLP in 1978, were issued by OECD in 1981 (Beernaert et al., 2008). Recently, OECD published a monograph on applying the principles of GLP to in vitro studies (Beernaert et al., 2008), which is an update of a topic addressed as early as 1971 by OECD. OECD has also published numerous guidelines for the testing of chemicals (e.g., Test Guideline 453: Combined Chronic Toxicity/Carcinogenicity Studies; OECD, 2009a), which was first published in 1981 and updated in 2009 (OECD, 2009b).

International cooperation was fostered between the U.S. and IARC (Weisburger, 1983). For example, US EPA published a list of carcinogens in 1980 (US EPA, 1980) that listed all the chemicals for which substantial or strong evidence existed showing that exposure to these chemicals either caused cancer in humans, or caused cancer in animals, which in turn made them potentially carcinogenic to humans, as per US EPA's 1976 Interim Guidelines for Carcinogenic Risk Assessment. Candidates for the list were either chemicals that US EPA's Cancer Assessment Group (CAG) had previously determined posed a potential human cancer risk or chemicals that one or more of three organizations—IARC, NTP, and FDA—had designated as potential human carcinogens.

Cancer classification schemes also became harmonized internationally. In 1980, Griesemer and Cueto from NCI reported that they had used the IARC Working Group's approach to evaluate the evidence for carcinogenicity from 198 NCI bioassays published from 1976 to 1980 (Griesemer and Cueto, 1980). They noted that an IARC working group had designated two categories of chemicals from the IARC Monographs with evidence of carcinogenicity in experimental animals: evidence that was "sufficient" or "limited." In attempting to apply this approach to the NCI bioassays, they developed a four-step process (including an evaluation of the quality of each study) and five categories of carcinogenicity: (1) very strong evidence in two species; (2) very strong evidence in one species and sufficient evidence in a second species; (3) very strong evidence in one species and no evidence in a second species; (4) sufficient evidence in two species; and (5) sufficient evidence in one species and no evidence in a second species. In Supplement 7, IARC subsequently developed four categories (sufficient evidence, limited evidence, inadequate evidence and lack of carcinogenicity [IARC, 1987, pp. 30, 31] and later five categories still used today: carcinogenic to humans, probably or possibly carcinogenic to humans, not classifiable, and not carcinogenic to humans [IARC, 1991, pp. 27–28]). US EPA essentially adopted these same categories (US EPA, 1986a).

International interdisciplinary teams and discussion groups became the standard method of resolving the various issues that arose from the use of animal studies for characterizing human cancer risks. The International Life Sciences Institute (ILSI) sponsored an Interdisciplinary Discussion Group on Carcinogenicity studies held in Chapel Hill, North Carolina, which brought together biologists and statisticians from around the world to address a number of issues such as biological and statistical assumptions in the analysis of the bioassay, dose selection, criteria for classifying neoplasms and use of non-neoplastic lesions, challenges to experimental design, and interpretation of multiple studies and design of repeat studies (Grice and Ciminera, 1988). At this meeting, NTP scientists presented their recently published guidelines for combining neoplasms for evaluation of rodent carcinogenesis studies, stressing the need to evaluate malignant and benign tumors both separately and in combination and opining that the most appropriate use of the animal tumor data in quantitative risk assessment must be evaluated on a case-by-case basis (McConnell et al., 1986).

The appropriate use of data in risk assessment (and in setting standards) also became a major focus of institutions and scientists. The risk assessment approach was codified in the National Research Council's (NRC) *Risk Assessment in the Federal Government: Managing the Process* (NRC, 1983). US EPA formed the Risk Assessment Forum (RAF) to promote consensus on risk assessment issues, as defined in the 1983 NRC report, and to ensure the incorporation of this consensus into appropriate

agency risk assessment guidance. To fulfill this purpose, the RAF assembles risk assessment experts, who focus on issues fundamental to the risk assessment process and related science policy issues. The RAF has issued a number of publications starting in 1986 (US EPA, 2009), which have included guidelines on mutagenicity risk assessment (US EPA, 1986b), developmental toxicity risk assessment (US EPA, 1991a), the relevance of certain male rat kidney tumors for human risk assessment (US EPA, 1991b), reproductive toxicity risk assessment (US EPA, 1996), and harmonization in interspecies extrapolation (US EPA, 2006).

In addition, the RAF published the first non-interim Guidelines for Carcinogen Risk Assessment in 1985 (interim guidelines had been published by US EPA in 1976 [Albert, 1994]). These guidelines, which have been updated periodically in 1996, 1999, 2003, and most recently in 2005, set out how data, especially animal data, should be used in risk assessments.

The 1980s also saw the beginning of an effort to better understand the cellular and molecular steps of carcinogenesis as it applies to particular chemicals, in order to make more informed extrapolations of animal data to humans. Thus for example, papers by Anderson et al. (1980) and Dietz et al. (1981) suggested that pharmacokinetic and adduct formation data could be used in addition to bioassay results to improve risk assessment for VC. Subsequently, other researchers called for the use of mechanistic data more generally in dose-response assessment and risk assessment (Starr, 1985; Gibson and Starr, 1988; Conolly et al., 1988; Legator and Ward, 1991; Reed, 1991; Lutz, 1991; Portier, 1993; Vainio et al., 1992, Vainio, 1994). Much of the work related to incorporating mechanistic data into risk assessment was conducted at the Chemical Industry Institute of Toxicology (CIIT; now The Hamner Institutes for Health Sciences).

In particular, CIIT scientists conducted extensive work in the 1980s and 1990s to explore how mechanistic and pharmacokinetic data could be used to supplement the results of the chronic bioassay in risk assessment (Starr, 1985; Swenberg et al., 1987; Gibson and Starr, 1988). Most notable were the large number of publications concerning formaldehyde carcinogenesis (Swenberg et al., 1980; Kerns et al., 1983; Starr and Buck, 1984; Starr and Gibson, 1985; Connolly et al., 1988; Casanova et al., 1991; Brenneman et al., 2000). Data from rodent bioassays alone (where the primary finding concerned nasal tumors) suggested a substantial and surprising human cancer risk at environmental exposure levels. The dosimetry and mechanism of action data obtained and published by these scientists did much to help reconcile the animal data in terms of potential risk for humans (Brenneman et al., 2000).

Starting in the late 1980s, a number of scientists also investigated the possibility of using in vitro or short-term in vivo data as a replacement for the chronic bioassay. Much of this work was initially conducted by scientists at IARC. These studies, which generally took the form of evaluating the ability of short-term tests to predict bioassay results (Bartsch and Tomatis, 1983; Ennever et al., 1987; Bartsch and Malaveille, 1990), generally concluded that in vitro methods (e.g., various genotoxicity assays, structure activity relationships) were not sufficiently predictive to replace the chronic assay, although they were useful for chemical prioritization. Similar findings were published nearly two decades later in an evaluation of NTP data (Benigni and Zito, 2004). Thus, the chances of finding a faster and less expensive alternative to the chronic bioassay seemed low. However, recent advances in genomic science have allowed the identification of early gene expression changes in response to short-term carcinogen exposure that appear to be sufficiently predictive to substantially reduce if not eliminate the evaluation of chemicals via chronic bioassays (Thomas et al., 2007, 2009; Nioi et al., 2008; Hoenerhoff et al., 2009).

1990s: Understanding mechanisms and risk

In the 1990s, the testing of chemicals for carcinogenicity became even more closely entwined with quantitative risk assessment. It was no longer sufficient to know whether or not a chemical was carcinogenic; it was also important to answer questions about the mechanism of action, relative potency, and linearity of the dose-response curve. A key focus for toxicologists in this decade was non-genotoxic carcinogenesis, including the realization that it was an oversimplification to divide carcinogens into two simple categories—genotoxicants and non-genotoxicants (Butterworth, 1990). Studies suggested that non-genotoxic mechanisms of carcinogenesis could include cell killing with increased compensatory cell proliferation, proliferative responses to chronic inflammation, decreased cell-cell communication controlling cell replication, induction of activating enzymes, mimicry of endogenous hormones, and immunosuppression (Williams et al., 1996). A key question that followed from the greater recognition of non-genotoxic carcinogenesis was the possibility of thresholds for tumor formation. These factors all served to complicate carcinogenicity testing, requiring collection and analysis of additional physiological parameters (e.g., labeling indices) or placement of doses in order to bracket the presumed threshold. In practice it would be impossible to design a single bioassay to address all of these concerns and thus it became more common for the 2-year rodent bioassay to be accompanied by auxiliary studies attempting to address these issues.

Another topic of discussion in the 1990s was the issue of dose selection in the chronic bioassay. As environmental policy became increasingly based on avoiding risks of very low level exposures, and regulatory risk assessments were based on the results of animal studies conducted at much higher doses, the relevance of bioassay data came into question. The significance of effects observed at or near the maximum tolerated dose (MTD), in particular, was vigorously discussed (Ames and Gold, 1990; Gold

et al., 1998; Fung et al., 1995; Foran, 1997; Rall, 2000); this discussion continues today.

Another important development was the rapid growth in genomic science and the development of transgenic animals. Considerable effort was made to use specific transgenic rodent models (in particular a knock out mouse model hemizygous for p53, and another mouse model carrying an activated H-*ras* oncogene) as more sensitive test species (Tennant et al., 1996; Eastin, 1998; Gulezian et al., 2000). Being more sensitive, data on potential carcinogens (and, in particular, weak carcinogens) could be obtained more quickly and perhaps with fewer animals than in the traditional bioassay (Schwetz and Gaylor, 1997). Particularly in the pharmaceutical industry, where the focus is largely on preemptive hazard identification, these models have increasingly been used as substitutes for the traditional rodent bioassay (Long et al., 2010). Their adoption to the broader world of non-pharmaceutical testing, where the focus lies more heavily in dose-response assessment, has been much slower. One key difficulty lies in interpretation—if extrapolations between normal mice and humans are uncertain, extrapolations between humans and mice genetically enhanced to develop tumors must be even more uncertain.

Finally, we note that in 1997, the International Conference on Harmonisation (ICH) adopted guidelines for carcinogenicity testing in pharmaceuticals (ICH, 1997). Although the chronic components of the guidelines were largely consistent with the standard NTP protocol, the guidelines also suggested that alternative testing procedures (e.g., transgenic mouse models) could be more useful than a second standard bioassay in another species (Cohen, 2001).

2000 to the present: New priorities and approaches

As a further step in the development of the bioassay, scientists are now increasingly focused on cases in which animal testing has *not* been predictive for humans (e.g., chloroform and melamine carcinogenicity, rodent thyroid carcinogenesis from perchlorate and phenobarbitol), as well as on differences in the mode of action and pharmacokinetics between species and as a function of dose (Cohen et al., 2004; Meek et al., 2003; Holsapple et al., 2006). In response to such discoveries, US EPA′s most recent carcinogen risk assessment guidance (published in 2005) allowed for replacement of the default linear no-threshold approach with a consideration of a chemical′s mode of action (MOA) in determining both the relevance to humans and the appropriate low-dose extrapolation approach (US EPA, 2005). In considering species differences in MOA, scientists have determined that a number of tumor types observed in rodents do not occur in humans and are therefore unlikely to be carcinogenic in humans (e.g., thyroid follicular cell tumors in rats, $\alpha_2$uglobulin–mediated renal tumors in male rats, urinary bladder neoplasia in rats) (Hayes, 2008; Meek et al., 2003; Baetcke et al., 1991). The species specificity of certain tumor types has been known to researchers for quite some time, but it often takes several decades before the MOA becomes sufficiently understood to incorporate that information into decision making. For example, the bladder carcinogenicity of saccharin was first established in the 1970s, but it was not until many decades later that mechanistic studies established the findings as not relevant to humans, allowing the continued use of saccharin (NCI, 2007b).

Another important topic is what stocks and strains of animals should be used in testing. An NTP workshop conducted in 2005 explored what animal models should be used in the NTP rodent cancer bioassay. The workshop participants concluded that the high background incidence of some types of tumors (e.g., testicular interstitial cell tumors and mononuclear cell leukemia in the F344/N rat) was an issue (King-Herbert and Thayer, 2006). Additional issues unique to the NTP F344/N rat colony (at Taconic Farms) included declining fertility, sporadic seizure activity, and chylothorax. The workshop report strongly advised that the NTP discontinue use of this strain and either reestablish the F344/N from another source (recognizing that would not address the elevated background cancer rates), create an F1 hybrid, or consider using an alternative such as the outbred Wistar-Han rat. For mice, although the group did not recommend changes in the current mouse model, they did recognize the high incidence of background liver tumors and made suggestions should the NTP elect to explore the use of multiple mouse strains (King-Herbert and Thayer, 2006).

Public concerns regarding the extensive use of animals in research have also led many scientists to question whether the chronic animal bioassay is necessary in all cases or if alternatives that reduce or eliminate the need for experimental animals are sufficient (Goodman, 2001; Cohen, 2001; Gruber and Hartung, 2004; Höfer et al., 2004, Backland et al., 2005). NRC, in the recent publication, *Toxicity Testing In The Twenty-First Century: A Vision And A Strategy*, suggested ways to streamline chemical toxicity testing and focus more on mechanisms of toxicity that can be evaluated in vitro (NRC, 2007). As its name implies, the NRC document describes a process for transitioning away from "a system based on whole-animal testing to one founded primarily on in vitro methods that evaluate changes in biologic processes using cells, cell lines, or cellular components, primarily of human origin" (NRC, 2007). This suggestion is not so much a rejection of the bioassay as it is a recognition that animal testing, as presently designed and interpreted, is not capable of providing the robust and voluminous amount of data required to address current public health concerns.

Mechanisms of carcinogenesis are now better understood to the point where some agencies (e.g., NTP) have been conducting fewer long-term carcinogenicity bioassays and replacing them with mechanistic

studies (Cogliano, 2006). Using this approach, chemical compounds may be suspected of being potentially carcinogenic even though they have never been tested in a long-term bioassay in experimental animals. An IARC workshop on the use of short- and medium-term tests for carcinogens concluded that, in the absence of carcinogenicity bioassays in experimental animals, strong mechanistic data could be used to evaluate potential carcinogenicity (Cogliano, 2006).

A recent NRC publication, *Science and Decisions: Advancing Risk Assessment* (NRC, 2009), suggests harmonizing non-cancer and cancer dose-response approaches in risk assessment, in part by abandoning the current default approaches and substituting them with more scientifically derived dose-response functions. This could involve constructing dose-response relationships based on information from a variety of study types, including cancer bioassays and in vitro studies, augmented by mechanistic information such as pharmacokinetic factors (NRC, 2009, pp. 129, 135). In addition the NRC recommended a more careful consideration of background exposures and biologic susceptibility factors, which as noted by NRC differ substantially between animals and humans (NRC, 2009, p. 135). These changes—and others recommended by NRC—would likely continue to shift cancer testing in animals from the standard bioassays used for virtually all chemicals to the use of more innovative testing aimed at answering specific questions about how a particular chemical interacts with both animal and human biology.

Key methodological challenges in early cancer bioassay design

Consistent with the expansion of the cancer bioassay to address different types of chemicals, and the need for more standardized testing methodology, a shift in several key aspects of the bioassay occurred over time, specifically in the types of chemicals studied, the exposure duration, the exposure route used, and the animal species used. This can readily be seen by looking at details of study designs reported by the US Public Health Service (US PHS). The US PHS's *Survey of Compounds Which Have Been Tested for Carcinogenic Activity* (US PHS Publication Number 149) is a compendium of published animal studies that provide data on carcinogenicity. The Survey uniquely provides information on the design and results of tens of thousands of bioassays published over the past 100 years. It should be noted that the Survey was developed as a data compilation tool and took results as they were published by the researchers without screening them against any particular inclusion criteria. It thus casts a very broad and non-selective net in terms of the studies that were included. The first volume of the Survey was prepared by US PHS scientist Jonathan Hartwell in 1941 and covered studies up to 1939. A second edition was published in 1951 with two supplements following in 1957 and 1969. US PHS published subsequent volumes roughly every other year thereafter up to 2000 (US PHS, 1941–2000). The level of effort required to compile such a database in the 1940s and 1950s is hard to imagine today when studies are indexed electronically and many publications can be downloaded online.

The US PHS survey provides a unique window, not only into the results of animal studies being conducted during the formative period of bioassay development, but also into the types of studies used to evaluate potential carcinogenicity. Using this resource, we created a database of the information reported in the Survey by compiling records for studies published in the last 2 years of each decade (i.e., 1938–1939, 1948–1949, 1958–1959, 1968–1969, 1978–1979), the final period coinciding with the establishment of NTP in the late 1970s and the widespread use of the 2-year, two-species bioassay protocol. Due to the time and effort required to review the thousands of study entries and enter the appropriate data into a database, we elected to use the last 2 years of each decade as a frame of reference. Focusing on the last 2 years of the decade allowed us to avoid for the most part the period of World War II, which presumably would represent an aberrant period in toxicological research. Once the data were tabulated, they were analyzed to understand how bioassays evolved over time in terms of the experimental design employed (i.e., types of chemicals tested, study duration, dose route, and animal species).

The Survey summarizes experimental data by compound (i.e., all studies in a particular period providing results for compound "x"). Publications reporting data for more than one compound may therefore be listed more than once. In our analysis, we chose to treat individual chemical entries as individual data points. For a given entry, if multiple dose routes or different species were used, each route and/or species was counted. For a given entry, the study duration was characterized by the maximum duration in any experimental group.

We determined the number of total entries for each of four chemical classes described in the Survey (i.e., inorganics, aliphatics, monocyclics, bicyclics), as well as the number of entries that involved experiments of more than 1 year in duration. We then evaluated the data to determine the methodological features used in the studies and subsequently excluded some data from our analysis. An examination of the Survey data indicated that the polycyclic aromatic hydrocarbons (PAHs) were studied almost exclusively via dermal or subcutaneous exposures (and almost overwhelmingly in mice or occasionally rats) to understand structure-activity relationships. Including these data in our methodological analysis would have resulted in strong sample heterogeneity. In addition, the polycyclic, and to a lesser extent heterocyclic, chemical data sets were observed to more commonly contain results from a few individual studies that evaluated large numbers of compounds using the same study design. Such studies would introduce a bias, suggesting a more standardized study methodology than really existed. This study predominance was seen to occur to a much

lesser extent for the other compounds. We also chose not to include data for the steroids, because our primary interest is industrial or commercial chemicals and the steroids (primarily forms of estradiol, progesterone, and testosterone) would not fall into this class.

Within this subset of Survey data we examined studies for several study design variables: the route by which the dose was administered, the species tested, and the duration of the study. Information was retrieved from the relevant volume of the Survey, coded, and entered into a spreadsheet (Excel). Various sorting and counting functions were then used to characterize these study design variables over the timeframe of interest.

Types of chemicals evaluated

Not surprisingly, the types of chemicals being evaluated in the studies reviewed by US PHS shifted over time in concert with advances in chemistry and industrialization. As shown in Table 1, in the 1938–1939 period, the potential carcinogenicity of inorganics, aliphatics, monoaromatics, and bicyclics were all equally being studied, as suggested by the published research at this time. If one examines all of the studies published prior to 1947 (data not shown), a similar equality in distribution is seen. Subsequent to this period, aliphatic and monoaromatic organic compounds became increasingly important (Zapp and Doull, 1994).

Study duration

One area of discussion in the early literature concerned the period over which dosing should occur. The definition of "long-term" varied with respect to both the duration of dosing and the duration of the period of follow-up evaluation. In the 1950s and 1960s, conducting large-scale tests in laboratory animals was relatively new, and the scientific literature reflects ongoing discussion regarding a number of problems with the testing. Some are listed by Hartwell (1951) and by Shubik and Hartwell (1957), who caution that their survey did not provide the means for determining the likelihood that a compound was actually carcinogenic, because "many of these studies [were] not carried on for a long enough period, they often [did] not record pathological information in sufficient detail, and the bulk of these studies [were] confined to feeding experiments with rats, often strains with such spontaneous tumor incidence as to make the detection of carcinogenicity quite impossible." Over time, however, the study duration increased.

The US PHS authors explicitly included only those tests lasting more than 30 days, which they apparently considered to be the minimum duration for evaluating potential carcinogenicity. Other scientific commentators of the time recommended studies of longer duration. As described by Boyland, studies "should be at least 12 weeks and can be one year or the duration of the life of the animal" (Boyland, 1958). Lehman et al. (1955) recommended 12 to 30 months for a long-term study. Nonetheless, the testing was often conducted for less than one year. In Shubik and Hartwell's (1957) catalog of compounds tested for carcinogenic activity, fewer than one third of the animal tests listed had a duration of 1 year or more.

Over time, however, bioassays were conducted for longer periods of time, as demonstrated in Table 2. In the 1930s and 1940s, carcinogenicity studies of longer than 1 year were rare. In 1938–1939, for example, 88% of the studies were for less than 1 year and only 5% were greater than or equal to 2 years. By 1958–1959, 18% of studies listed in the Survey had been conducted for 2 years or more, and by 1978–1979, this number had increased to 25%. The number of studies greater than or equal to 1 year in duration in 1978–1979 was still only 54%, although a number of the short-term studies in this timeframe appeared to be focused on studying cellular or molecular mechanisms of toxicity (e.g., examining DNA adduct formation), effects where short study times are necessary to detect the first signs of cellular damage. Thus, although the duration of assays designed to test carcinogenicity increased over time, there was also an increased use of short-term studies to investigate carcinogenic mechanisms in chemicals already known to be mutagenic or carcinogenic.

The period over which an assay is conducted has remained a subject of discussion. For example, Mauderly et al. (1987) exposed rats to diesel exhaust via inhalation for up to 30 months and found that of all the rats with tumors, only 19% were identified at or before the 24-month termination and 81% were identified later. They noted that the concern that an increasing control incidence of tumors might render a cancer study insensitive to treatment-related tumors beyond 24 months was not borne out. For decades, the Ramazzini Foundation has routinely conducted cancer bioassays in which the animals are typically evaluated at the end of their full lifespan rather than at two years, as under the protocol of many agencies worldwide (e.g., NTP, OECD). This difference is important because the incidence of background tumors rises substantially at the end of the rodent's lifespan, creating additional "noise" in the data and complicating the interpretation of the results. The appropriateness of conducting bioassays in this manner has been debated in the scientific literature, most notably in the context of assessing the potential carcinogenicity of the artificial sweetener aspartame (Soffriti et al., 2007; EFSA, 2006).

Proponents of full-lifetime studies aver that such studies detect chemicals that might selectively increase tumor incidence at old ages, which the 2-year study design fails to address. Proponents of the 2-year study design argue that the natural increase in tumors after 2 years of age in rodents makes it difficult to statistically separate treatment-related effects from the normal increase in tumor incidence with aging, leading to an increase in false positive interpretations. In addition, concerns have been raised that when animals die spontaneously and necropsy is delayed, tissues may become autolyzed, impeding accurate diagnosis. This debate remains one

Table 1. Classes of chemicals studied in cancer bioassays as reported in the US PHS serial *Survey of Compounds Which Have Been Tested for Carcinogenic Activity*.

	1938–1939	1948–1949	1958–1959	1968–1969	1978–1979
Inorganics	4	4	20	21	27
Aliphatics	3	8	20	91	148
Monoaromatics	3	17	18	107	138
Bicyclics	2	3	2	22	30

Note. Data were obtained for the last 2 years of each decade from the 1930s to the 1970s.

Table 2. Design features of cancer bioassays as reported in the US PHS serial *Survey of Compounds Which Have Been Tested for Carcinogenic Activity*.

	Percent of entries				
Percent of entries	1938–1939	1948–1949	1958–1959	1968–1969	1978–1979
By study duration:					
<1 year	88	79	55	42	49
≥1 year	7	18	45	58	54
≥2 years	5	7	18	11	25
By route of exposure:					
Oral	22	62	31	58	76
Inhalation	0	0	5	2	2
Dermal	17	6	8	9	11
Subcutaneous	28	26	28	18	13
Implantation	22	2	11	5	4
Intraperitoneal	0	2	7	8	9
Other	22	6	0	7	22
By test species:					
Rats	11	51	55	40	62
Mice	39	30	35	50	47
Rabbits	33	2	4	3	1
Guinea pigs	0	2	4	1	1
Hamsters	0	2	5	4	8
Dogs	22	17	8	3	4
Primates	0	2	1	2	1
Other	0	4	0	2	2

Note. Data were obtained for the last 2 years of each decade from the 1930s to the 1970s.

of the more contentious topics in the area of cancer risk assessment (see for example, Soffritti et al., 2007; EFSA, 2006).

Dose route

As shown in Table 2, over time, the routes of exposure used in animal studies steadily shifted from routes of exposure not as relevant to humans, such as subcutaneous injection and implantation, to oral dosing. Exposure via implantation, although initially common, was little used by 1978–1979. Dermal testing, as a percentage of tests having at least a 1-year duration, stayed relatively similar from 1938–1939 to 1978–1979.[2] Inhalation testing, although absent in the 1938–1939 and 1948–1949 study groups, remained at about 2% to 5% of animal studies from 1958–1959, 1968–1969, and 1978–1979. Oral dosing, which was used with similar frequency to subcutaneous injection and implantation in 1938–1939, has predominated since 1948–1949.

Dose

A critical issue in bioassay design that continues to be debated is the question of dose selection (Counts and Goodman, 1995; Rhomberg et al., 2007). If the goal of the cancer bioassay is purely qualitative (i.e., is the chemical a carcinogen or not?), then only one dose besides zero is required and all the available animals can be dosed at this level to maximize study sensitivity. If, on the other hand, a goal of the bioassay is to provide some understanding of how carcinogenic response changes with dose, then as many doses as possible are required and the number of animals per dose must be reduced accordingly. Early in the development of the cancer bioassay it was decided to include a maximum possible dose that would not jeopardize the animals′ long-term survival (FDA, 1959) or at least cause no more than 10% to 20% to succumb to long-term toxicity (Weisburger and Weisburger, 1967). The rationale was that if a chemical was a weak carcinogen, then giving the maximum possible dose would increase the chance the carcinogenic effect would be detected in the limited number of animals used. This MTD thus became a required component of the bioassay. Over time, various ways to specify the characteristics of the MTD have been proposed; currently it is described as the dose that produces no shortening of the animals' lives or

no more than a 10% decrease in body weight (Haseman, 1985).

Use of tumor data collected at the MTD began to be questioned in the late 1970s (Munro, 1977; Roe, 1980; Haseman, 1985; Ames and Gold, 1990). A number of scientists opined that the host of biological reactions that occur at the MTD may not be relevant for low-dose carcinogenesis in humans (Roe, 1980; Ames and Gold, 1990; Gaylor, 2005). For example, dosing at the MTD may overwhelm the body′s defense capabilities (i.e., deplete reserves of vital antioxidants) or cause tissue damage that stimulates reparative cell proliferation. An analysis by Gaylor in 2005 of NTP historical bioassay data suggested that tumor occurrence at the MTD may be a general effect of high-dose damage (e.g., cell killing and compensatory proliferation) rather than any chemical-specific property (e.g., specific DNA adduct formation).

In addition, data collected at the MTD are, at best, of limited use for understanding effects at low environmental levels. As bioassays have increasingly been seen as opportunities to develop dose-response information as opposed to only providing hazard identification data (e.g., in recent suggestions of using the benchmark dose approach for carcinogens), inclusion of the MTD in studies may provide data of less value for risk assessment than if an additional dose were placed at lower concentrations. A recent ILSI monograph has suggested that more detailed consideration of the goals for the study (i.e., qualitative hazard identification, quantitative dose-response analysis) should occur prior to selection of the dosing regimen (Rhomberg et al., 2007).

Animal species

One significant variable in bioassay testing prior to the late 1960s was the wide variety of animals used by different researchers. Shubik and Hartwell (1957) demonstrated that scientists were using a surprising array of species in animal tests: mice, rabbits, guinea pigs, rats, pigs, dogs, cats, salamanders, various non-human primates, hamsters, goats, chickens, horses, calves, ducks, etc., listing 22 in all. Shubik and Hartwell (1969) reported that 26 different species had been used for chronic toxicity testing, not including dozens of different strains of mice and rats.

Over time, however, rats and mice became the predominant animal species used in studies of at least one-year duration, as shown in Table 2. In 1938–1939, mice and rabbits were the most frequently used animals, whereas rats were only used in about 11% of the studies. However, by the late 1940s, the predominance of rats and mice in chronic bioassays was largely established. The use of dogs has decreased consistently over time (since 1938–1939), whereas the use of non-human primates remained constant at about 1 to 2% for 1948–1949, 1958–1959, 1968–1969, and 1978–1979.

A key issue that fostered the adoption of mice and rats as standard test species and allowed for longer-duration studies was the development of specific pathogen–free (SPF) animals in the 1960s (Lane-Petter, 1962; Bell et al., 1964; Saquet, 1965). Such animals, delivered by cesarean section under aseptic conditions and then maintained in highly controlled environments, are free of common and highly communicable rodent pathogens (a fact that can be confirmed by routine testing before the start of an experiment). Prior to the advent of SPF animals, long-term studies were at risk of disease outbreaks (e.g., murine pneumonia) that, if not fatal to the entire colony, nevertheless complicated interpretation of the study results and left the study′s validity in question (Lane-Petter, 1962, 1970; Bell et al., 1964).

More recently, respiratory infections in test animals have been at the center of a controversy regarding results of certain animal tests conducted at the Ramazzini Foundation in Italy, where a team of scientists from the NTP found greater occurrence of lung inflammation and lower frequency of leukemia and lymphoma of the lung as compared to that reported by the Ramazzini Foundation (Malarkey et al., 2010). As a result of these developments, US EPA has suspended its ongoing risk assessment process for four chemicals where the Ramazzini Foundation′s findings were pivotal to the evaluation of carcinogenicity (US EPA, 2010).

Limitations

Our findings derived from the US PHS database have limitations. In analyzing the US PHS database, we relied on the authors′ interpretations of what constituted an appropriate study design for evaluating carcinogenicity. Clearly, studies considered valid in the past are very different from studies considered valid today. Because our objective was to examine the state of the science from the 1930s through the 1970s, the period when the chronic animal bioassay was undergoing its most fundamental development, we used the opinions expressed in the period in which the studies were conducted. We also made another assumption in using the US PHS data: we used each chemical-specific study reported by US PHS as our primary source of information and counted it as one data point, and we did not adjust for instances in which published studies reported results for multiple chemicals (studies that would be listed multiple times in the US PHS reports). Consequently, our results may overestimate the number of publications related to animal testing in a given period, but we believe the effect to be fairly limited because over the entire period we evaluated, the majority of published studies were focused on a single chemical.

In addition to the US PHS database of carcinogenicity studies, we are also aware of the carcinogenic potency database maintained by the University of California at Berkeley (http://potency.berkeley.edu). This online database (in effect, a continuation of the survey) is an extremely useful tool for reviewing comprehensive carcinogenicity test data for a large number of chemicals. The database extends only back to the 1950s and thus would not allow evaluation of animal bioassays from the 1930s

and 1940s (UCB and Gold, 2008). In addition, studies included in the University of California database are prescreened with a set of quality criteria relating to factors such as animal number, study duration (e.g., greater than 6 months for rodents), etc. These criteria, which reflect contemporary attitudes on data quality standards, would eliminate many studies of historical interest and would not allow us to fully evaluate past historical practice. The data provided in the US PHS studies were more comprehensive for the entire period we evaluated.

Conclusions

To give a sense of historical opinion on the role and validity of toxicological testing, we reviewed scientific articles to obtain a sense of the majority opinion among scientific commentators at various points in time. Clearly, there will always be differences of opinion among scientists about any issue as fundamental to the field of toxicology as the validity of animal models. Likewise, during any period there will be scientists whose calls for action will later be shown to be foresighted, but whose recommendations will be slowly adopted, if at all. The writings of such individuals, although important, do not reflect the actual state of the science at the time (indeed, the tone of some of these writings suggests the state of the science was lagging far behind).

From primarily a research focus aimed at studying known effects in humans of exposure to chemical carcinogens, the goal of the bioassay gradually evolved towards predicting the toxicity of agents in foods, drugs, and subsequently environmental and occupational chemicals. From the 1940s through the early 1970s, these tests were also used primarily for what we would now call hazard identification (i.e., whether the chemical produces the effect). Starting in the late 1970s, bioassays were increasingly used to gather quantitative dose-response data on which to base acceptable exposure levels. More recently the objectives of testing have become broader, with test goals encompassing information regarding mode of action and other parameters aimed at supporting risk assessment and evaluating the relevance of results to humans.

Test methods and study designs have also evolved to become more robust and more specialized. It was not until the late 1960s that standardized protocols involving near-lifetime exposures of mice and rats and standardized pathology batteries came into broad use. The wide adoption of a standard approach in study design improved the validity of these types of studies for predicting potential human health risks and developing appropriate public health regulation. More recently, testing strategies emphasize a suite of tests including multiple in vitro tests, short-term animal tests, and long-term animal bioassays. Current debates focus on the needs for test methods to address chemical-specific modes/mechanisms of action (e.g., endocrine action), and the need to limit the number of animals used in research. Efforts are also ongoing to develop ways of refining the 2-year bioassay, for example, by initially conducting genomics screens and toxicokinetic/toxicodynamic studies that can help to optimize study design (e.g., by identifying the inflection points in the dose-response curve when dose-dependent mechanisms of action are believed to be important). Thus, the use of chronic testing in toxicology is a continuously evolving process and defining a "standard approach" is only possible within the context of a specific time period.

It now appears that the field is about to enter a new stage of development, with a shift away from animal-intensive study methods towards a greater emphasis on in vitro or even in silico techniques. The goal of this shift should be to develop new test protocols that are more predictive of human health risks and not just adequate replacements for chronic animal tests, adding to the uncertainty in extrapolation. There is a real opportunity for developing such tests. As with earlier periods, the shift to the new testing protocols will be gradual, the new tests gaining from the interplay between our ability to understand the significance of the resulting data and our ability to improve test design. In a sense, this is a dialectic process in which testing protocols and strategies (and their scientific underpinnings) are presented, rejected, and sometimes resurrected, until scientists resolve the issues at hand and in so doing, transform science.

Declaration of interest

Some of the initial research describe in this manuscript was funded in conjunction with a litigation matter where one of the authors (B. D. Beck) served as an expert witness. However, the majority of the research and the preparation of the manuscript was supported by the individual authors and their employers. The opinions expressed are solely those of the authors.

References

Albert RE. (1994). Carcinogen risk assessment in the U.S. Environmental Protection Agency. Crit Rev Toxicol 24:75–85.

Ames BN, Gold LS. (1990). Chemical carcinogenesis: Too many rodent carcinogens. Proc Natl Acad Sci U S A 87:7772–7776.

Anderson MW, Hoel DG, Kaplan NL. (1980). A general scheme for the incorporation of pharmacokinetics in low-dose risk estimation for chemical carcinogenesis: Example - —Vinyl chloride. Toxicol Appl Pharmacol 55:154–161.

Backland S, Winder C, Khalil C, Hayes A. (2005). Toxicity assessment of industrial chemicals and airborne contaminants: Transition from *in vivo* to *in vitro* test methods: A review. Inhal Toxicol 17:775–787.

Baetcke KP, Hard GC, Rodgers IS, McGaughy RE, Tahan LM. (1991). Alpha2u-globulin: Association with Chemically Induced Renal Toxicity and Neoplasia in the Male Rat. Report to US Environmental Protection Agency, Washington, DC: Risk Assessment Forum. Report no. EPA/625/3-91/019F. September 1991.

Ballentine C. (1981). Taste of raspberries, taste of death: The 1937 elixir sulfanilamide incident. FDA Consumer Magazine. June.

Bartsch H, Malaveille C. (1990). Screening assays for carcinogenic agents and mixtures: An appraisal based on data in the IARC monograph series. IARC Sci Publ 104:65–74.

Bartsch H, Tomatis L. (1983). Comparison between carcinogenicity and mutagenicity based on chemicals evaluated in the IARC monographs. Environ Health Perspect 47:305–317.

Beernaert H, Vanherle AM, Bertrand S. (2008). Critical aspects in implementing the OECD monograph No. 14 ""The application of the principles of GLP to in vitro studies."." *Ann Ist Super Sanita* 44:348–356.

Bell DP, Elmes PC, Wheeler SM. (1964). A colony of specific pathogen-free rats. Nature 201:273–274.

Benigni R, Zito R. (2004). The second National Toxicology Program comparative exercise on the prediction of rodent carcinogenicity: Definitive results. Mutat Res 566:49–63.

Boorman GA, Montgomery CA, Eustis SL, Wolf MJ, McConnell EE, Hardisty JF (1985). "Quality assurance in pathology for rodent carcinogenicity studies." In: Milman HA, Weisburger EK, eds. Handbook of Carcinogen Testing. HA Milman,and EK Weisburger (eds). Noyes Publications, Park Ridge, NJ.: Noyes Publications, pp. 345-–357.

Boyland E. (1958). The biological examination of carcinogenic substances. Br Med Bull 14:93–98.

Brenneman KA, Conolly RB, Gaido KW, Greenlee WF, Kimbell JS, Recio L. (2000). Research at CIIT on the health effects of exposure to chemicals. CIIT Activities 20:9–10.

Breslow L, Wilner D, Agran L, Breslow DL, Ellwein LB, Morganstern M, eds. (1978). A History of Cancer Control in the United States 1946–1971: A History of Scientific and Technical Advances in Cancer Control. Bethesda, MD: US Dept.artment of Health, Education, and Welfare, National Cancer Institute.

Butterworth BE. (1990). Consideration of both genotoxic and nongenotoxic mechanisms in predicting carcinogenic potential. Mutat Res 239:117–132.

Carson R. (1962). Silent Spring. Boston: Houghton Mifflin.

Casanova M, Morgan KT, Steinhagen WH, Everitt JI, Popp JA, Heck HD. (1991). Covalent binding of inhaled formaldehyde to DNA in the respiratory tract of rhesus monkeys: Pharmacokinetics, rat-to-monkey interspecies scaling, and extrapolation to man. Fundam Appl Toxicol 17:409–428.

Case RAM, Hosker ME. (1954). Tumour of the urinary bladder as an occupational disease in the rubber industry in England and Wales. Br J Prev Soc Med 8:39–50.

Cogliano, VJ. 2006. Use of carcinogenicity bioassays in the IARC monographs. Ann. N Y Acad. Sci. 1076:592–600.

Cohen SM. (2001). Alternative models for carcinogenicity testing: Weight of evidence evaluations across models. Toxicol Pathol 29(Suppl):183–190.

Cohen SM, Klaunig J, Meek ME, Hill RN, Pastoor T, Lehman-McKeeman L, Bucher J, Longfellow DG, Seed J, Dellarco V, Fenner-Crisp P, Patton D. (2004). Evaluating the human relevance of chemically-induced animal tumors. Toxicol Sci 78:181–186.

Conolly RB, Reitz RH, Clewell HJ 3rd, Andersen ME. (1988). Pharmacokinetics, biochemical mechanism and mutation accumulation: A comprehensive model of chemical carcinogenesis. Toxicol Lett 43:189–200.

Counts JL, Goodman JI. (1995). Principles underlying dose selection for, and extrapolation from, the carcinogen bioassay: Dose influences mechanism. Regul Toxicol Pharmacol 21:418–421.

Creech JL, Johnson MN. (1974). Angiosarcoma of liver in the manufacture of polyvinyl chloride. J Occup Med 16:150–151.

Dietz FK, Reitz RH, Watanabe PG, Gehring PJ. (1981). Translation of pharmacokinetic/biochemical data into risk assessment. Adv Exp Med Biol 136(Pt B):1399–1424.

Eastin WC. (1998). The U.S. National Toxicology Program evaluation of transgenic mice as predictive models for identifying carcinogens. Environ Health Perspect 106(Suppl 1):81–4.

EFSA. (2006). Opinion of the Scientific Panel on Food Additives, Flavourings, Processing Aids and Materials in Contact with Food (AFC) on a request from the Commission related to a new long-term carcinogenicity study on aspartame: Question number EFSA-Q-2005-122. Parma, Italy: European Food Safety Authority.

Ennever FK, Noonan TJ, Rosenkranz HS. (1987). The predictivity of animal bioassays and short-term genotoxicity tests for carcinogenicity and non-carcinogenicity to humans. Mutagenesis 2:73–78.

Fairhall LT. (1952). Industrial toxicology: Dinitro-ortho-cresol. Occup Health (Auckl) 12:132–133.

FDA. (1959). Appraisal of the Safety of Chemicals in Foods, Drugs, and Cosmetics. Washington, DC: US Food and Drug Administration. Association of Food and Drug Officials of the United States.

FDA. (1978). Nonclinical laboratory studies: Good laboratory practice regulations. Fed Regist 43:59986-–60025. US Food and Drug Administration. 21 CFR Part 58, December 22.

FDA. (2005). Milestones in US Food and Drug Law History. US Food and Drug Administration. Department of Health and Human Services. Available at: http://www.fda.gov/opacom/backgrounders/miles.html.

FDA. (2009). NCTR: A Look Back. US Food and Drug Administration. Available at: http://www.fda.gov/AboutFDA/WhatWeDo/History/FOrgsHistory/NCTR/ucm080414.htm.

FDA. (2010). Tolerances for Poisonous Ingredients in Food. Sec. 406. [21 USC § 346]. US Food and Drug Administration. Available at: http://www.fda.gov/RegulatoryInformation/Legislation/FederalFoodDrugandCosmeticActFDCAct/FDCActChapterIVFood/ucm107545.

Fitzhugh OG, Nelson AA. (1946). Comparison of the chronic toxicity of triethylene glycol with that of diethylene glycol. J Ind Hyg Toxicol 28: 40–43.

Foran JA. (1997). Principles for the Selection of Doses in Chronic Rodent Bioassays. Washington, DC: ILSI Press.

Frederick WG. (1984). The birth of the ACGIH Threshold Limit Values Committee and its influence on the development of industrial hygiene. In: LaNier ME, ed. Annals of the American Conference of Governmental Industrial Hygienists,. Vol. 9. Cincinnati, OH: American Conference of Governmental Industrial Hygienists (ACGIH), 11-–13.

Fung VA, Barrett JC, Huff J. (1995). The carcinogenesis bioassay in perspective: Application in identifying human cancer hazards. Environ Health Perspect 103:680–683.

Gaylor DW. (1980). The ED01 study: Summary and conclusions. J Environ Pathol Toxicol 3(3 Spec No):179–183.

Gaylor DW. (2005). Are tumor incidence rates from chronic bioassays telling us what we need to know about carcinogens? Regul Toxicol Pharmacol 41:128–133.

Gaylor DW, Aylward LL. (2004). An evaluation of benchmark dose methodology for non-cancer continuous-data health effects in animals due to exposures to dioxin (TCDD). Regul Toxicol Pharmacol 40:9–17.

Gibson JE, Starr TB. (1988). Opportunities for improving techniques for interspecies extrapolation in the risk assessment process. Environ Health Perspect 77:99–105.

Gold LS, Slone TH, Ames BN. (1998). What do animal cancer tests tell us about human cancer risk?: Overview of analyses of the Carcinogenic Potency Database. Drug Metab Rev 30:359–404.

Goodman JI. (2001). A perspective on current and future uses of alternative models for carcinogenicity testing. Toxicol Pathol 29(Suppl):173–176.

Grice HC, Ciminera JL, eds. (1988). Carcinogenicity: The Design, Analysis and Interpretation of Long-Term Animal Studies. Report to International Life Sciences Institute (ILSI). New York: Springer-Verlag.

Griesemer RA, Cueto C. (1980). Toward a classification scheme for degrees of experimental evidence for the carcinogenicity of chemicals for animals. In: Montesano R, Bartsch H, Tomatis L, eds. Molecular and Cellular Aspects of Carcinogen Screening Tests. IARC Scientific Publications No. 27. Lyon, France: International Agency for Research on Cancer, 259-–281.

Gruber FP, Hartung T. (2004). Alternatives to animal experimentation in basic research. ALTEX 1(Suppl 1):3–31.

Gulezian D, Jacobson-Kram D, McCullough CB, Olson H, Recio L, Robinson D, Storer R, Tennant R, Ward JM, Neumann DA.

(2000). Use of transgenic animals for carcinogenicity testing: Considerations and implications for risk assessment. Toxicol Pathol 28:482–499.

Hackmann C. (1958). Problems of testing preparations for carcinogenic properties in the chemical industry. In: O′Connor O′,Connor M, ed. CIBA Foundation Symposium on Carcinogenesis Mechanisms of Action, Wolstenholme, GEW. Boston: Little, Brown and Co., 308–322.

Hadidian Z, Fredrickson TN, Weisburger EK, Weisburger JH, Glass RM, Mantel N. (1968). Tests for chemical carcinogens. Report on the activity of derivatives of aromatic amines, nitrosamines, quinolines, nitroalkanes, amides, epoxides, aziridines, and purine antimetabolites. J Natl Cancer Inst 41:985–1036.

Hall WE. (1951). The industrial hazards of the organic phosphates and related insecticides, including parathion. Ann West Med Surg 5:1025–1027.

Hartwell JL. (1951). Survey of Compounds which have been Tested for Carcinogenic Activity. (Seco2nd edition). ([excerpts).]. Bethesda, MD: National Cancer Institute, 1–4, 546, 570.

Haseman JK. (1985). Issues in carcinogenicity testing: Dose selection. Fundam Appl Toxicol 5:66–78.

Hayes, AW, ed. (2008). Principles and Methods of Toxicology. (Fif5th edition). Philadelphia, PA: Taylor & Francis.

Hoenerhoff MJ, Hong HH, Ton TV, Lahousse SA, Sills RC. (2009). A review of the molecular mechanisms of chemically induced neoplasia in rat and mouse models in National Toxicology Program bioassays and their relevance to human cancer. Toxicol Pathol 37:835–848.

Höfer T, Gerner I, Gundert-Remy U, Liebsch M, Schulte A, Spielmann H, Vogel R, Wettig K. (2004). Animal testing and alternative approaches for the human health risk assessment under the proposed new European chemicals regulation. Arch Toxicol 78:549–564.

Holsapple MP, Pitot H, Cohen SH, Boobis AR, Klaunig JE, Pastoor T, Dellarco VL, Dragan YP. (2006). Mode of action in relevance of rodent liver tumors to human cancer risk. Toxicol Sci 89:51–56.

Hueper WC. (1957). Newer developments in occupational and environmental cancer. Arch Intern Med 100:487–503.

Hueper WC, Wiley FH, Wolfe HD, Ranta KE, Leming MF, Blood FR. (1938). Experimental production of bladder tumors in dogs by administration of betanaphthylamine. J Ind Hyg Toxicol 20:46–84.

IARC. (1972). IARC Monographs on the Evaluation of Carcinogenic Risk of Chemicals to Man,. Vol. 1. Lyon: International Agency for Research on Cancer., World Health Organization.

IARC. (1979). IARC Monographs on the Evaluation of Carcinogenic Risk of Chemicals to Humans: Vol. 20. Some Halogenated Hydrocarbons,. Vol 20. Lyon, International Agency for Research on Cancer., World Health Organization.

IARC. (1979). IARC Monographs on the Evaluation of Carcinogenic Risk of Chemicals to Humans: Some Halogenated Hydrocarbons, Volume 20. World Health Organization (WHO).

IARC. (1980). IARC Monographs, Supplement 2: Long-Term and Short-Term Screening Assays for Carcinogens: A Critical Appraisal. Lyon: International Agency for Research on Cancer., World Health Organization.

IARC. (1987). IARC Monographs on the Evaluation of Carcinogenic Risks to Humans: Overall Evaluations of Carcinogenicity: An Updating of IARC Monographs, Volumes 1 to 42, Supplement 7. Lyon: International Agency for Research on Cancer., World Health Organization.

IARC. (1991). IARC Monographs on the Evaluation of Carcinogenic Risk of Chemicals to Humans: Vol. 53. Occupational Exposures in Insecticide Application, and Some Pesticides. Lyon: International Agency for Research on Cancer., World Health Organization.

ICH. (1997). S1B: Testing for Carcinogenicity of Pharmaceuticals. International Conference on HarmonizationHarmonisation. Available at: http://www.ich.org/LOB/media/MEDIA490.pdf.

Innes JRM, Ulland BM, Valerio MG, Petrucelli L, Fishbein L, Hart ER, Palotta AH, Bates RR, Falk HL, Gart JJ, Klein M, Mitchell I, Peters J. (1969). Bioassay of pesticides and industrial chemicals for tumorigenicity in mice: A preliminary note. J Natl Cancer Inst 42:1101–1115.

JECFA. (1958). Procedures for the Testing of Intentional Food Additives to Establish their Safety for Use: Second Report of the Joint FAO/WHO Expert Committee on Food Additives. Rome: Joint FAO/WHO Expert Committee on Food Additives., World Health Organization. WHO Technical Report Series No. 144.

Junod S. (1999). The rise and fall of federal food standards in the United States: The case of the peanut butter and jelly sandwich. Presented at Society for the Social History of Medicine., Spring conference, Aberdeen, Scotland. April 9.

Kalberer JT, Newell GR. (1979). Funding impact of the National Cancer Act and beyond. Cancer Res 39:4274–4284.

Keplinger ML, Goode JW, Gordon DE, Calandra JC. (1975). Interim results of exposure of rats, hamsters, and mice to vinyl chloride. Ann N Y Acad Sci 246:219–224.

Kerns WD, Pavkov KL, Donofrio DJ, Gralla EJ, Swenberg JA. (1983). Carcinogenicity of formaldehyde in rats and mice after long-term inhalation exposure. Cancer Res 43:4382–4392.

King-Herbert A; Thayer K. (2006). NTP workshop: Animal models for the NTP rodent cancer bioassay: Stocks and strains-–—Should we switch? Toxicol.Pathol. 34:802-–805.

Lane-Petter W. (1962). The provision and use of pathogen-free animals. Proc R Soc Med 55:253–256.

Lane-Petter W. (1970). A ventilation barrier to the spread of infection in laboratory animal colonies. Lab Anim 4:125–134.

Legator MS, Ward JB Jr. (1991). Use of *in vivo* genetic toxicity data for risk assessment. Mutat Res 250:457–65.

Lehman AJ. (1951). Proof of safety: Some interpretations. J Am Pharm Assoc 11:305–308.

Lehman AJ, Laug EP, Woodard G, Draize JH, Fitzhugh OG, Nelson AA. (1949). Procedures for the appraisal of the toxicity of chemicals in foods. Food Drug Cosmet Q 4:412–433.

Lehman AJ, Patterson WI, Davidow B, Hagan EC, Woodard G, Laug EP, Frawley JP, Fitzhugh OG, Bourke AR, Draize JH, Nelson AA, Vos BJ. (1955). Procedures for the appraisal of the toxicity of chemicals in foods, drugs, and cosmetics. Food Drug Cosmet Law J 10:679–748.

Long GG, Morton D, Peters T, Short B, Skydsgaard M. (2010). Alternative mouse models for carcinogenicity assessment: Industry use and issues with pathology interpretation. Toxicol Pathol 38:43–50.

Lutz WK. (1991). Dose-response relationships in chemical carcinogenesis: From DNA adducts to tumor incidence. Adv Exp Med Biol 283:151–156.

Malarkey D, Herbert R, Nyska A, Sutphin M, Pernicka K. [National Toxicology Program (NTP)]. (2010). Internal correspondence to J. Bucher (NTP) re: report on visit (4/25/2010 -–4/30/2010) and assessment of the pathology procedures performed at the Ramazzini Institute, Bentivoglio, Italy. June 11.

Maltoni C, Lefemine G. (1974a). [Experiments testing the carcinogenic potential of vinyl chloride.] Rendiconti (Accademia nazionale dei Lincei. Classe di scienze fisiche, matematiche e naturali). 56:1–15.

Maltoni C, Lefemine G. (1974b). Carcinogenicity bioassays of vinyl chloride. I. Research plan and early results. Environ Res 7:387–405.

Maltoni C, Lefemine G. (1975). Carcinogenicity bioassays of vinyl chloride. Current results. Ann N Y Acad Sci 246:195–218.

Mantel N, Bryan WR. (1961). Safety testing of carcinogenic agents. J Natl Cancer Inst 27:455–471.

Maron DM, Ames BN. (1983). Revised methods for the *Salmonella* mutagenicity test. Mutat. Res. 113 (3-4):173-–215.

Mauderly JL, Jones RK, Griffith WC, Henderson RF, McClellan RO. (1987). Diesel exhaust is a pulmonary carcinogen in rats exposed chronically by inhalation. Fundam Appl Toxicol 9:208–221.

McCann J, Choi E, Yamasaki E, Ames BN. (1975) Detection of carcinogens as mutagens in the Salmonella/microsome test: Assay of 300 chemicals. Proc. Natl. Acad. Sci. U S A 72 (12):5135–5139.

McConnell EE, Solleveld HA, Swenberg JA, Boorman GA. (1986). Guidelines for combining neoplasms for evaluation of rodent carcinogenesis studies. J Natl Cancer Inst 76:283–289.

Meek ME, Bucher JR, Cohen SM, Dellarco V, Hill RN, Lehman-McKeeman LD, Longfellow DG, Pastoor T, Seed J, Patton DE. (2003). A framework for human relevance analysis of information on carcinogenic modes of action. Crit Rev Toxicol 33:591–653.

Munro IC. (1977). Considerations in chronic toxicity testing: The chemical, the dose, the design. J Environ Pathol Toxicol 1:183–197.

Murray JA. (1922). The production of cancer by specific forms of irritation. Br Med J 2:1103–1104.

NCI. (2007a). Cancer Facts and the War on Cancer. National Cancer Institute. Division of Cancer Cause and Prevention. Available at: http://training.seer.cancer.gov/disease/war/.

NCI. (2007b). Artificial Sweeteners and Cancer. National Cancer Institute. Available at: www.cancer.gov/cancertopics/factsheet/Risk/artificial-sweeteners

Nelson AA, Fitzhugh OG, Calvery HO. (1943). Liver tumors following cirrhosis caused by selenium in rats. Cancer Res 3:230–236.

Nelson AA, Fitzhugh OG, Calvery HO. (1945). Production of bladder stones and bladder tumors in rats by feeding of diethylene glycol. Fed Proc Fed Am Soc Exp Pathol 4:149.

Nioi P, Pardo ID, Sherratt PJ, Snyder RD. (2008). Prediction of non-genotoxic carcinogenesis in rats using changes in gene expression following acute dosing. Chem Biol Interact 172:206–215.

NIOSH. (1996). NIOSH Celebrates 25th Anniversary. National Institute for Occupational Safety and Health. Available at: http://www.cdc.gov/Niosh/updates/twenfif.html. Accessed on 16 December 2008.

NRC. (1983). Risk Assessment in the Federal Government: Managing the Process. National Research Council. Committee on the Institutional Means for Assessment of Risks to Public Health. Washington, DC: National Academies Press.

NRC. (2007). Toxicity Testing in the Twenty-First Century: A Vision and a Strategy. National Research Council. Division on Earth and Life Studies, Board on Environmental Studies and Toxicology and Institute for Laboratory Animal Research. Washington, DC: National Academies Press.

NRC. (2009). Science and Decisions: Advancing Risk Assessment. National Research Council. Committee on Improving Risk Analysis Approaches Used by the U.S. EPA. Washington, DC: National Academies Press.

NTP. (1984). Report of the NTP Ad Hoc Panel on Chemical Carcinogenesis Testing and Evaluation. Research Park Triangle, NC: National Toxicology Program.

NTP. (2004). History of the NTP. National Toxicology Program. Available at: http://ntp.niehs.nih.gov/?objectid=720163C9-BDB7-CEBA-FE4B970B9E72BF54.

NTP. (2006). Eleventh Report on Carcinogens. National Toxicology Program. US Department of Health and Human Services, Public Health Service. Available at: http://ntp.niehs.nih.gov/index.cfm?objectid=32BA9724-F1F6-975E-7FCE50709CB4C932.

OECD. (2009a). OECD Test Guideline 453: Combined Chronic Toxicity\/Carcinogenicity Studies. Organisation for Economic Co-operation and Development. OECD 453. Available at: http://www.oecdilibrary.org/oecd/content/book/9789264071223-en. Accessed on 16 December 2008.

OECD. (2009b). OECD Guidelines for Testing of Chemicals: Full List of Test Guidelines. Organisation for Economic Co-operation and Development. Available at: http://www.oecd.org/dataoecd/8/11/42451771.pdf.

Page NP. (1977). Chronic toxicity and carcinogenicity guidelines. J. Environ. Pathol. Toxicol 1(2):161–182.

Patty FA, Yant WP, Waite CP. (1930). Acute response of guinea pigs to vapors of some new commercial organic compounds. V. Vinyl chloride. Public Health Rep 45:1963–1971.

Peto R, Gray R, Brantom P, Grasso P. (1991). Effects on 4080 rats of chronic ingestion of *N*-nitrosodiethylamine or *N*-nitrosodimethylamine: A detailed dose-response study. Cancer Res 51:6415–6451.

Portier CJ. (1993). Mechanistic modeling and risk assessment. Pharmacol Toxicol 72(Suppl 1):28–32.

Rall DP. (1984). The National Toxicology Program. In: Tegeris AS, ed. Toxicology Laboratory Design and Management for the 80′s 80′s and Beyond. Concepts in Toxicology, Vol. 1. Basel, Switzerland: Karger, 283–293.

Rall DP. (2000). Laboratory animal tests and human cancer. Drug Metab Rev 32:119–128.

Reed DJ. (1991). Future research needs for the application of mechanistic data to risk assessment. Adv Exp Med Biol 283:863–868.

Rhomberg LR, Baetcke K, Blancato J, Bus J, Cohen S, Conolly R, Dixit R, Doe J, Ekelman K, Fenner-Crisp P, Harvey P, Hattis D, Jacobs A, Jacobson-Kram D, Lewandowski T, Liteplo R, Pelkonen O, Rice J, Somers D, Turturro A, West W, Olin S. (2007). Issues in the design and interpretation of chronic toxicity and carcinogenicity studies in rodents: Approaches to dose selection. Crit Rev Toxicol 37:729–837.

Rodricks JV. (1988). Origins of risk assessment in food safety decision making. J Am Coll Toxicol 7:539–542.

Roe FJ. (1980). Current methods in testing for carcinogenic activity. Br J Cancer. 41:495–496.

Rohwer SA, Haller HL, Dubois K, Grob D, Abrams HK, Hamblin DC, Marchand JF, Lehman AJ, Hartzell A, Ward JW. (1950). Pharmacology and toxicology of certain organic phosphorus insecticides; general description of their activity and usefulness; pharmacology; toxicology; clinical experience; effects on beneficial forms of life, crops and soil and residue hazards. J Am Med Assoc 144:104–108.

Sacquet E. (1965). Pathogen-free animals. Food Cosmet Toxicol 3:47–55.

Schwetz B, Gaylor D. (1997). New directions for predicting carcinogenesis. Mol Carcinog 20:275–279.

Selling L. (1916). Benzol as leucotoxin. Studies on the degeneration and regeneration of the blood and haematopoietic organs. Johns Hopkins Hosp Rep 17:83–149.

Shubik P, Hartwell JL. (1957). Survey of Compounds which have been Tested for Carcinogenic Activity: Supplement 1. Washington, DC: US Government Printing Office.

Shubik P, Hartwell JL. (1969). Survey of Compounds which have been Tested for Carcinogenic Activity: Supplement 2. Washington, DC: US Government Printing Office.

Soffriti M, Belpoggi F, Minardi F, Maltoni C. (2002). Ramazzini Foundation Cancer Program: History and major projects, life span carcinogenicity bioassay design, chemicals studies, and results. Ann N Y Acad Sci 982:26–45

Soffritti M, Belpoggi F, Tibaldi E, Esposti DD, Lauriola M. (2007) Life-span exposure to low doses of aspartame beginning during prenatal life increases cancer effects in rats. Environ Health Perspect 115(9):1293–1297.

Sontag JM, Page NP, Saffiotti U. (1976). Guidelines for Carcinogen Bioassay in Small Rodents. Report to US Dept.artment of Health, Education and Welfare. Springfield, VA: National Technical Information Service. Report no. NTIS PB-264061, NCI-CG-TR-1.

SOT. (1981). Re-examination of the ED01 study. Review of statistics: The need for realistic statistical models for risk assessment. Society of Toxicology: ED01 Task Force. Fundam Appl Toxicol 1:124–126.

Starr TB. (1985). The role of mechanistic data in dose-response modeling. Basic Life Sci 33:101–124.

Starr TB, Buck RD. (1984). The importance of delivered dose in estimating low-dose cancer risk from inhalation exposure to formaldehyde. Fundam Appl Toxicol 4:740–753.

Starr TB, Gibson JE. (1985). The mechanistic toxicology of formaldehyde and its implications for quantitative risk estimation. Annu Rev Pharmacol Toxicol 25:745–767.

Swenberg JA, Kerns WD, Mitchell RI, Gralla EJ, Pavkov KL. (1980). Induction of squamous cell carcinomas of the rat nasal cavity by inhalation exposure to formaldehyde vapor. Cancer Res 40(9):3398–3402.

Swenberg JA, Richardson FC, Boucheron JA, Deal FH, Belinsky SA, Charbonneau M, Short BG. (1987). High- to low-dose extrapolation: Critical determinants involved in the dose response of carcinogenic substances. Environ Health Perspect 76:57–63.

Tennant RW, Spalding J, French JE. (1996). Evaluation of transgenic mouse bioassays for identifying carcinogens and noncarcinogens. Mutat Res 365:119–127.

Thomas RS, Bao W, Chu TM, Bessarabova M, Nikolskaya T, Nikolsky Y, Andersen ME, Wolfinger RD. (2009). Use of short-term transcriptional profiles to assess the long-term cancer-related safety of environmental and industrial chemicals. Toxicol Sci 112:311–321.

Thomas RS, O′Connell O′Connell TM, Pluta L, Wolfinger RD, Yang L, Page TJ. (2007). A comparison of transcriptomic and metabonomic technologies for identifying biomarkers predictive of two-year rodent cancer bioassays. Toxicol Sci 96:40–46.

Tomatis L. (1979). The predictive value of rodent carcinogenicity tests in the evaluation of human risks. Annu Rev Pharmacol Toxicol 19:511–530.

Tomatis L. (1986). Foreword. In: Montesano R, Bartsch H, Vainio H, Wilbourn J, Yamasaki H, eds. Long-Term and Short-Term Assays for Carcinogens: A Critical Appraisal. International Agency for Research on Cancer (IARC) Scientific Publication No. 83, p1. Lyon: IARC.

Tsutsui H. (1918). "[Short summary of the original essay: On the artificial cancer produced in mice]. GANN 12:17–21.

UCB, Gold LS. (2008). The carcinogenic potency database (CPDB). University of California, Berkeley. Available at: http://potency.berkeley.edu/.

US EPA. (1976). Health risk and economic impact assessments of suspected carcinogens: Interim procedures and guidelines. Fed Regist 41(1102):21402–21405.

US EPA. (1980). The Carcinogen Assessment Group′s Group′s List of Carcinogens. Washington, DC: Carcinogen Assessment Group. Report no. EPA 450/R-80–104. July 1980.

US EPA. (1986a). Guidelines for Carcinogen Risk Assessment. Washington, DC: Carcinogen Assessment Group. Report no. EPA/630/R-00/004. September 1986.

US EPA. (1986b). Guidelines for Mutagenicity Risk Assessment. Washington, DC: Risk Assessment Forum. Report no. EPA/630/R-98/003. September 1986.

US EPA. (1986). Guidelines for Mutagenicity Risk Assessment. Washington, DC: Risk Assessment Forum. Report no. EPA/630/R-98/003. September 1986.

US EPA. (1991a). Guidelines for Developmental Toxicity Risk Assessment. Washington, DC: U.S. Environmental Protection Agency, Risk Assessment Forum, Washington, DC,. EPA/600/FR-91/001,. 1991.

US EPA. (1991b). Report of the EPA Peer Review Workshop on Alpha2u-Globulin: Association with Renal Toxicity and Neoplasia in the Male Rat. Washington, DC: Risk Assessment Forum. Report no. EPA/625/3-91/021.

US EPA. (1996). Guidelines for Reproductive Toxicity Risk Assessment. Washington, DC: Risk Assessment Forum. Report no. 630/R-96/009.

US EPA. (2005). Guidelines for Carcinogen Risk Assessment (Final). Washington, DC: Risk Assessment Forum. Report no. EPA/630/P-03/001F.

US EPA. (2006). Harmonization in Interspecies Extrapolation: Use of BW3/4 as Default Method in Derivation of the Oral RfD. External Review Draft Report no. EPA/630/R-06/001. February 2006.

US EPA. (2009). Risk Assessment Forum Publications List. Office of the Science Advisor. Available at: http://www.epa.gov/raf/pubyear.htm. Accessed on 7 June 2010.

US EPA. (2010). News Release: EPA Places Four IRIS Assessments on Hold Pending Review. Available at: http://yosemite.epa.gov/opa/admpress.nsf/0/B64D44F06A56D5B285257742007C5002. Accessed on 15 June 2010.

US PHS. (1941-–2000). Survey of Compounds Which Have Been Tested for Carcinogenic Activity. US Public Health Service Pub. lication No. 149. US Public Health Service, National Cancer Institute. Washington, DC: US Government Printing Office.

Vainio H. (1994). Use of mechanistic and other data in identifying carcinogens: A review based on the IARC Monographs programme. Arch Toxicol Suppl. 16:281–294.

Vainio H, Magee PN, McGregor DB, McMichael AJ, eds. (1992). Mechanisms of Carcinogenesis in Risk Identification. IARC Scientific Publications No. 116. Lyon, France: World Health Organization International Agency for Research on Cancer. Scientific Publications No. 116.

Ward JM, Goodman DG, Griesemer RA, Hardisty JF, Schueler RL, Squire RA, Strandberg JD. (1978) Quality assurance of pathology in rodent carcinogenesis tests. J. Environ Pathol Toxicol Oncol. 2:371-–378

Wax PM. (1995). Elixirs, diluents, and the passage of the 1938 Federal Food, Drug and Cosmetic Act. Ann Intern Med 122:456–461.

Weisburger EK. (1983). History of the bioassay program of the National Cancer Institute. Prog Exp Tumor Res 26:187–201.

Weisburger JH, Weisburger EK. (1967). Tests for chemical carcinogens. In: Busch H, ed. Methods in Cancer Research,. Volume Vol. I. New York: Academic Press, 307–398.

Weisburger JH. (1994). Does the Delaney clause of the U.S. food and drug laws prevent human cancers? Fund. am Appl. Tox. icol 22:483-–493.

Weisburger JH. (1999). Carcinogenicity and mutagenicity testing, then and now. Mutation Research 437:105–112.

Williams GM, Iatropoulos MJ, Weisburger JH. (1996). Chemical carcinogen mechanisms of action and implications for testing methodology. Exp Toxicol Pathol 48:101–111.

Woodard G, Calvery HO. (1943). Acute and chronic toxicity: Public health aspects. Ind Med 12:55–59.

Yamagiwa K, Ichikawa K. (1915). Ueber die kunstliche erzeugung von papillom (German) = [On the experimental induction of papillomas.] [German]. Verh Jpn Pathol Ges 5:142–149.

Yamagiwa K, Ichikawa K. (1918). Experimental study of the pathogenesis of carcinoma. J Cancer Res 3:1–29.

Zapp JA, Doull J. (1994). Industrial toxicology: Retrospect and prospect. In: Patty′s Patty′s Industrial Hygiene and Toxicology. Fourth 4th editioned., Volume Vol. II, Part A. New York: Wiley and Sons, 1–24.